StyleCLIP - Project Outline

Alex Leonardi, Ben Harpe, Michael Fein

AC295

AC215

Pavlos Protopapas
Institute for Applied Computational Science, Harvard

Outline

- Project Scope
- Project Workflow
- Process Flow
- Data
- Models

Problem Definition

Pictures of various people exist but are normally pretty boring and we have little control over them. If you want to edit pictures in exciting ways, you have to learn Photoshop, which is quite difficult. There is no easy way to apply transformations to people in images and we'd like to be able to do so without much time or computational power.

Proposed Solution

We'll have an app on our phone that allows us to take or upload pictures. We will then be able to type in a description of the current picture and a description of the transformation we'd like to make. The app will then make that transformation to the image and return it. For example, we could take a picture of Pavlos and caption it "a face." If we added the transformation description "a bald face," nothing would happen.

Project Scope

Proof Of Concept (POC)

- Be able to run pretrained StyleCLIP model on our images
- Find additional images to train on (Celeb_a_HQ dataset)
- Finetune e4e model (used for StyleCLIP global) on new images
- Connect new model to CLIP and be able to generate new transformations

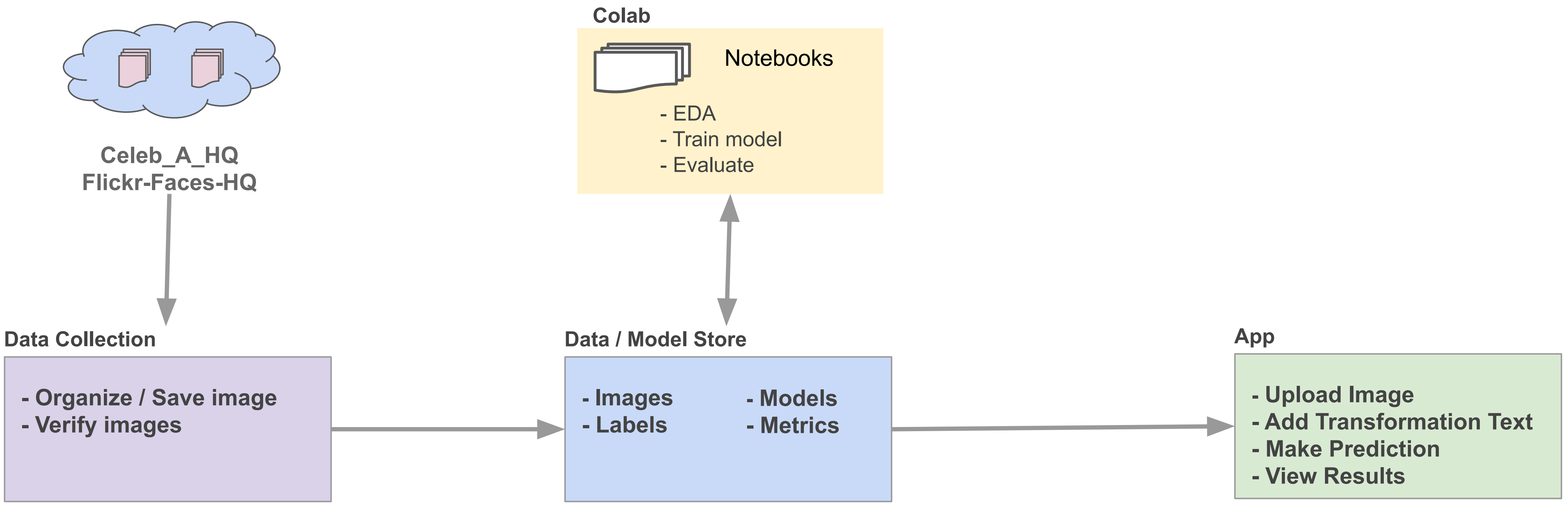
Prototype

- Create a mockup of screens to see how the app could look like
- Deploy StyleCLIP with finetuned e4e to Fast API to service model predictions as an API

Minimum Viable Product (MVP)

- Create containerized app to apply transformations to images
- API Server for uploading images and predicting using best model

Process Flow



Colab

Notebooks

- EDA
- Train model
- Evaluate

Celeb_A_HQ
Flickr-Faces-HQ

Data Collection

- Organize / Save image
- Verify images

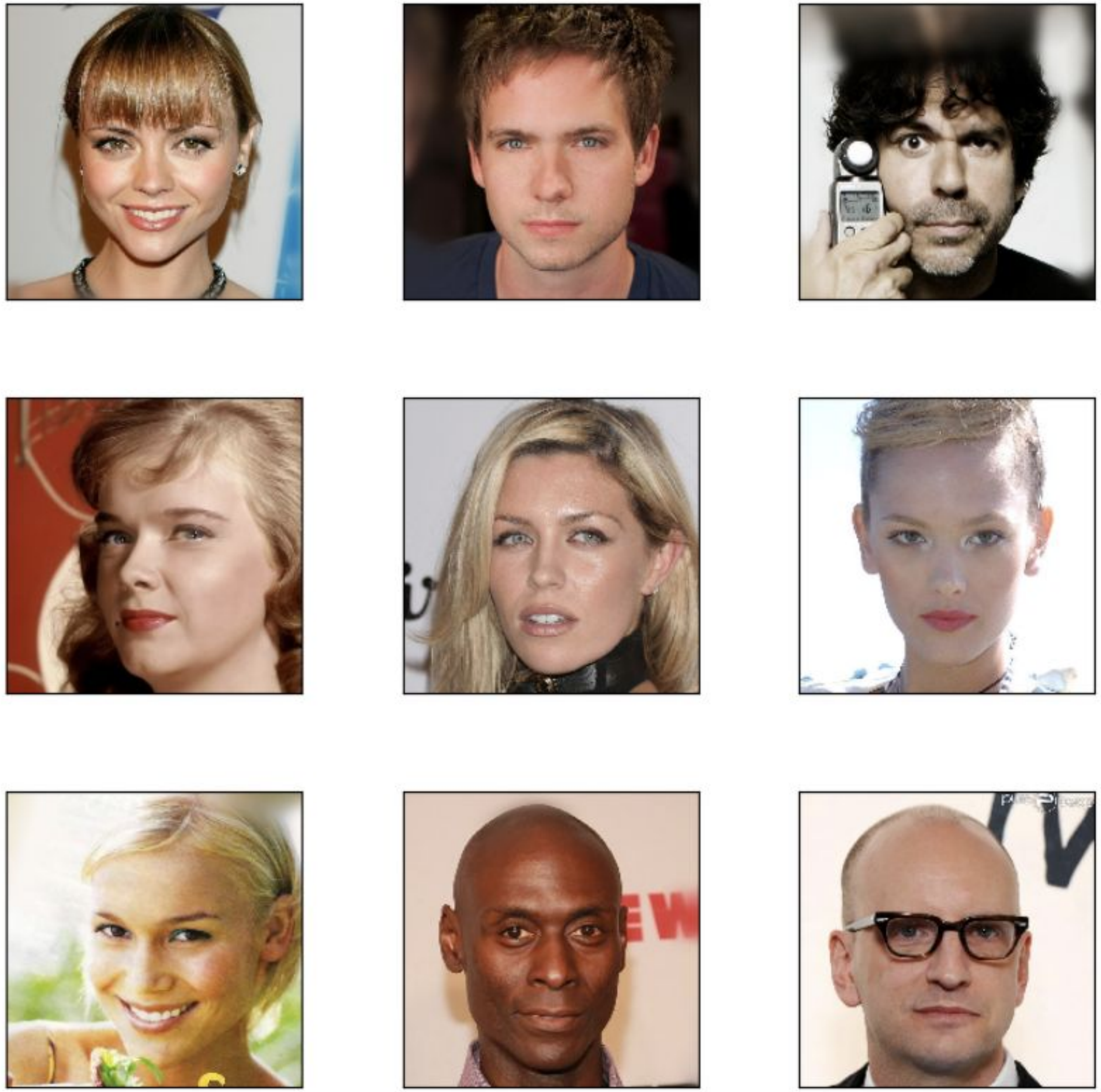
Data / Model Store

- Images
- Labels
- Models
- Metrics

App

- Upload Image
- Add Transformation Text
- Make Prediction
- View Results

Data - celeb_a_hq



Data Details

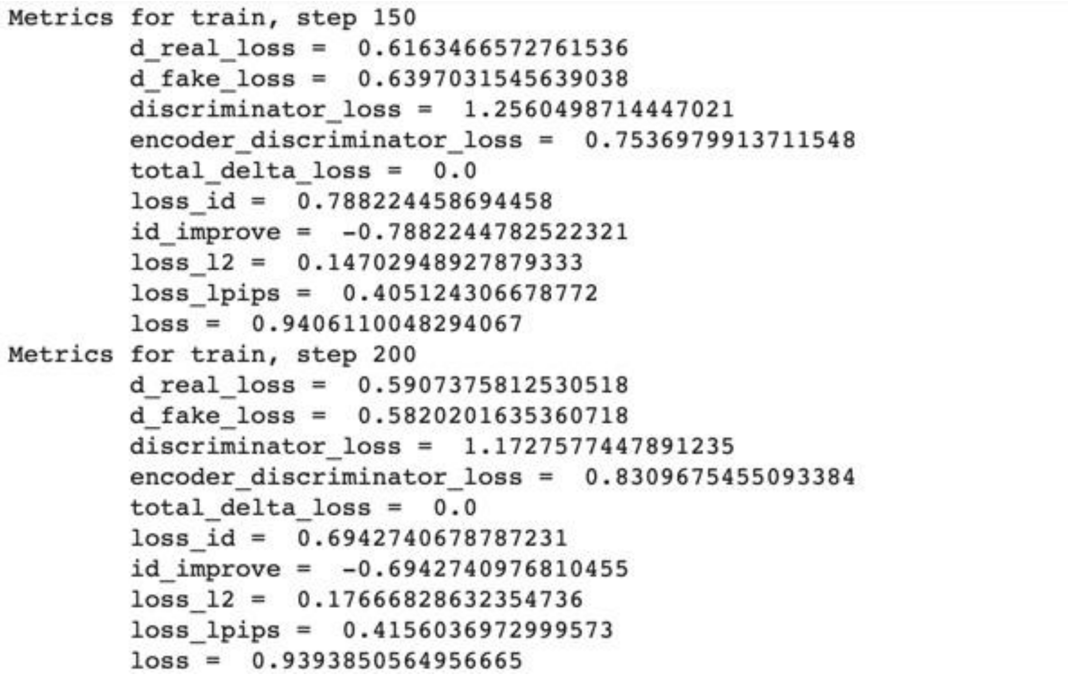
- Total number of images: 30000
- Dimensions: (1024,1024)
- Attributes: 40 binary, 5 landmarks

Models

Encoding of images:

- Transfer learning, e4e model

Models - Training Results



```
Metrics for train, step 150
        d_real_loss =  0.6163466572761536
        d_fake_loss =  0.6397031545639038
        discriminator_loss =  1.2560498714447021
        encoder_discriminator_loss =  0.7536979913711548
        total_delta_loss =  0.0
        loss_id =  0.788224458694458
        id_improve =  -0.7882244782522321
        loss_l2 =  0.14702948927879333
        loss_lpips =  0.405124306678772
        loss =  0.9406110048294067
Metrics for train, step 200
        d_real_loss =  0.5907375812530518
        d_fake_loss =  0.5820201635360718
        discriminator_loss =  1.1727577447891235
        encoder_discriminator_loss =  0.8309675455093384
        total_delta_loss =  0.0
        loss_id =  0.6942740678787231
        id_improve =  -0.6942740976810455
        loss_l2 =  0.17666828632354736
        loss_lpips =  0.4156036972999573
        loss =  0.9393850564956665
```